Doctor's Helper analysis

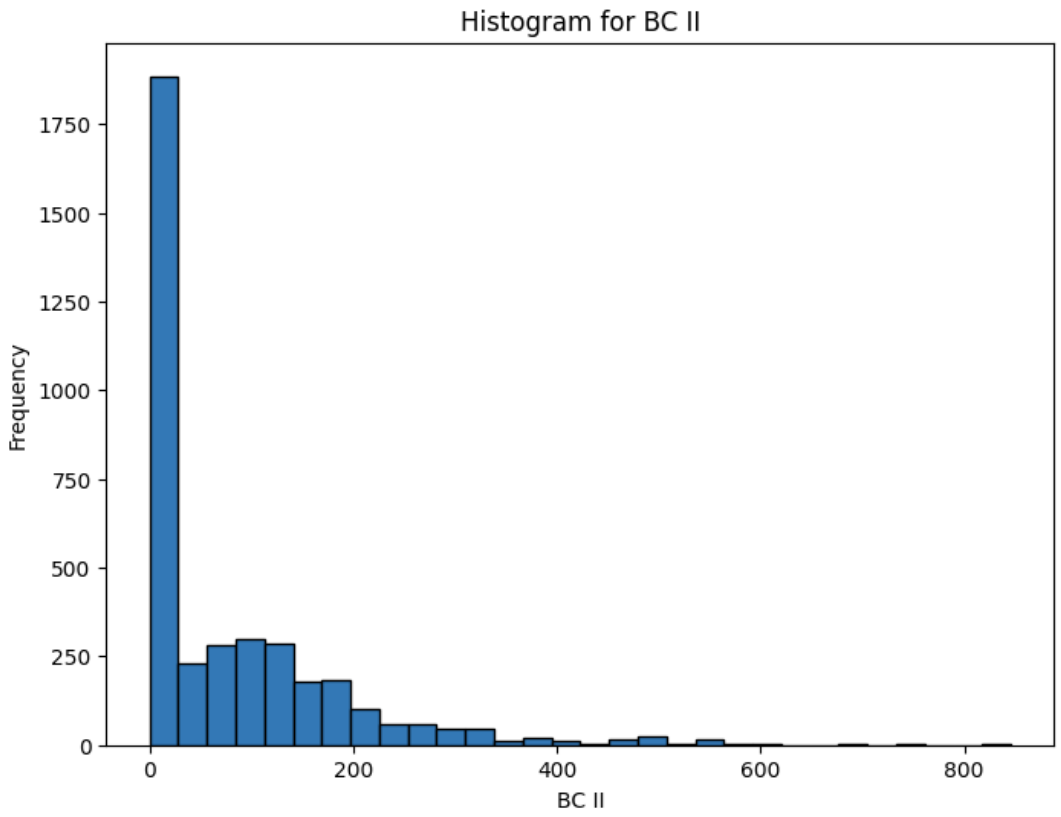
SZYMON DURAJ

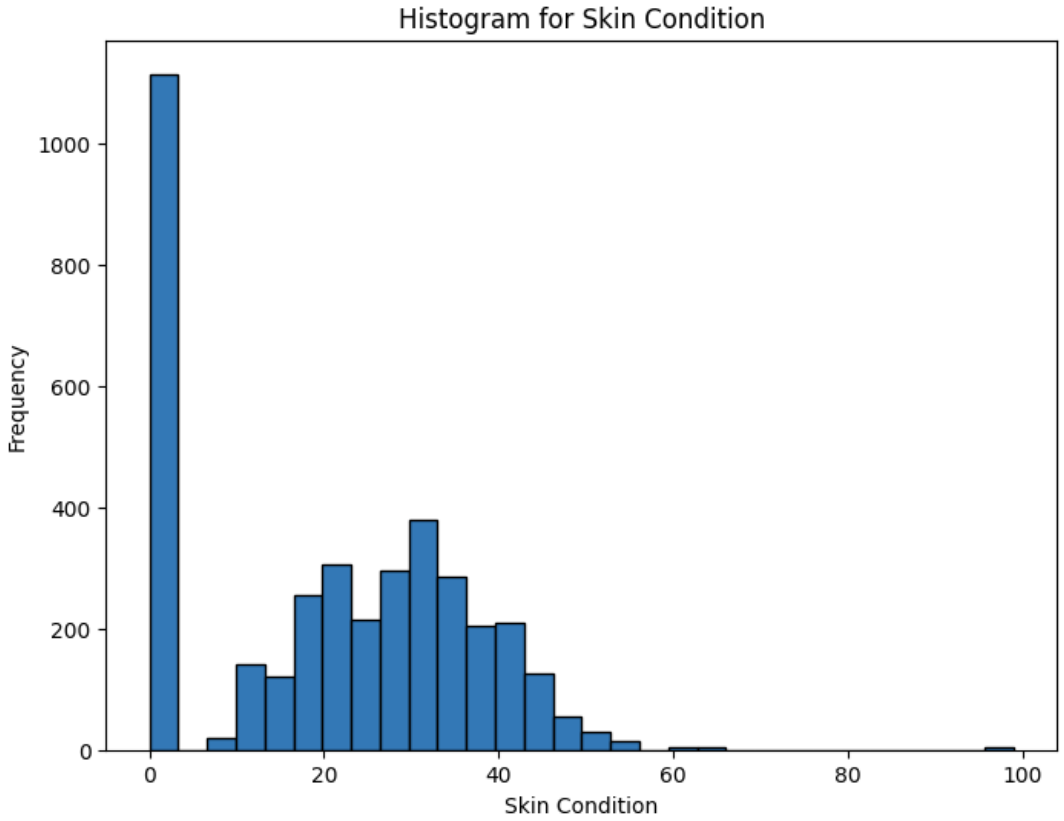
Let's take a look into our initial data:

	Unique_ID	# Pregnancies	Blood Chemestry~I	Blood Chemisty~II	Blood Chemisty~III	Blood Pressure	Skin Thickness	BMI	Genetic Predisposition Factor	Age	Air Qual'ty Index	$tate	Outcome
0	5642118.0	1.0	0.0	23.0	10.0	74.0	20.0	27.7	0.299	21.0	38.0	CA	0.0
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	5642119.0	1.0	0.0	0.0	61.0	68.0	35.0	32.0	0.389	22.0	10.0	CA	0.0
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	5642120.0	1.0	0.0	0.0	172.0	48.0	20.0	24.7	0.140	22.0	77.0	CA	0.0
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	5642121.0	5.0	44.0	0.0	207.0	62.0	0.0	25.0	0.587	36.0	40.0	OR	0.0
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	5642122.0	2.0	56.0	45.0	150.0	56.0	28.0	24.2	0.332	22.0	70.0	CT	0.0
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

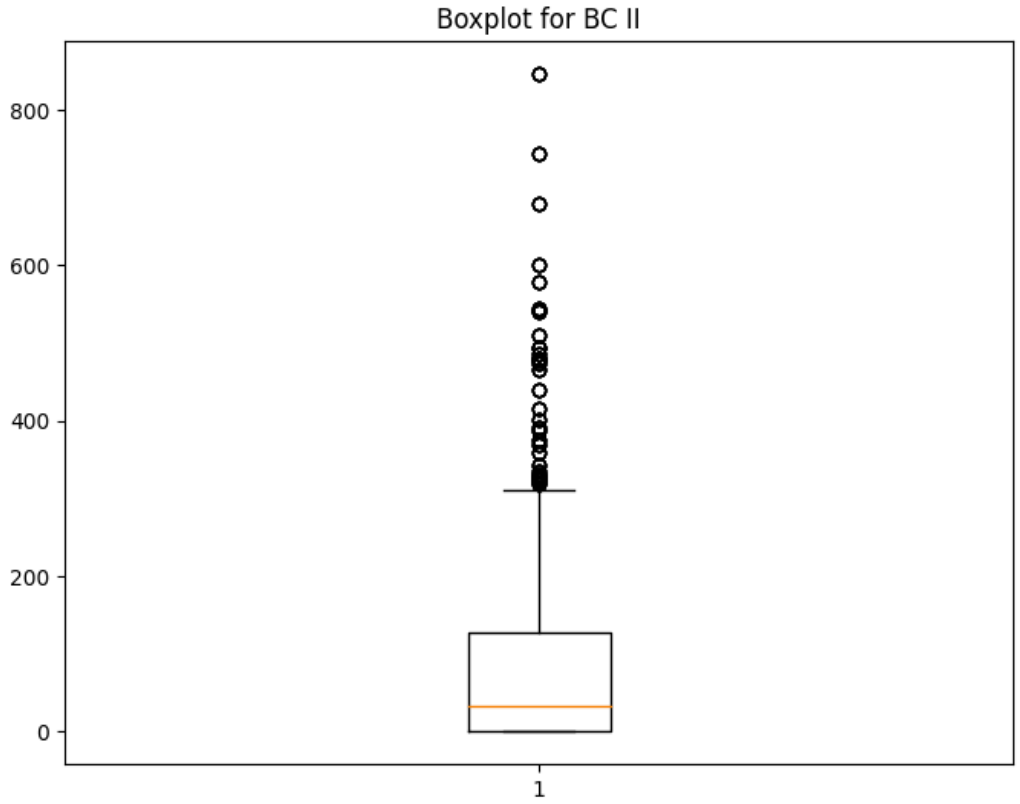
We can see that the data consist a lot of rows with NaN values and names of variables are incorrect comparing with the documentation which we received so we want to correct it and make the data quicker to use.

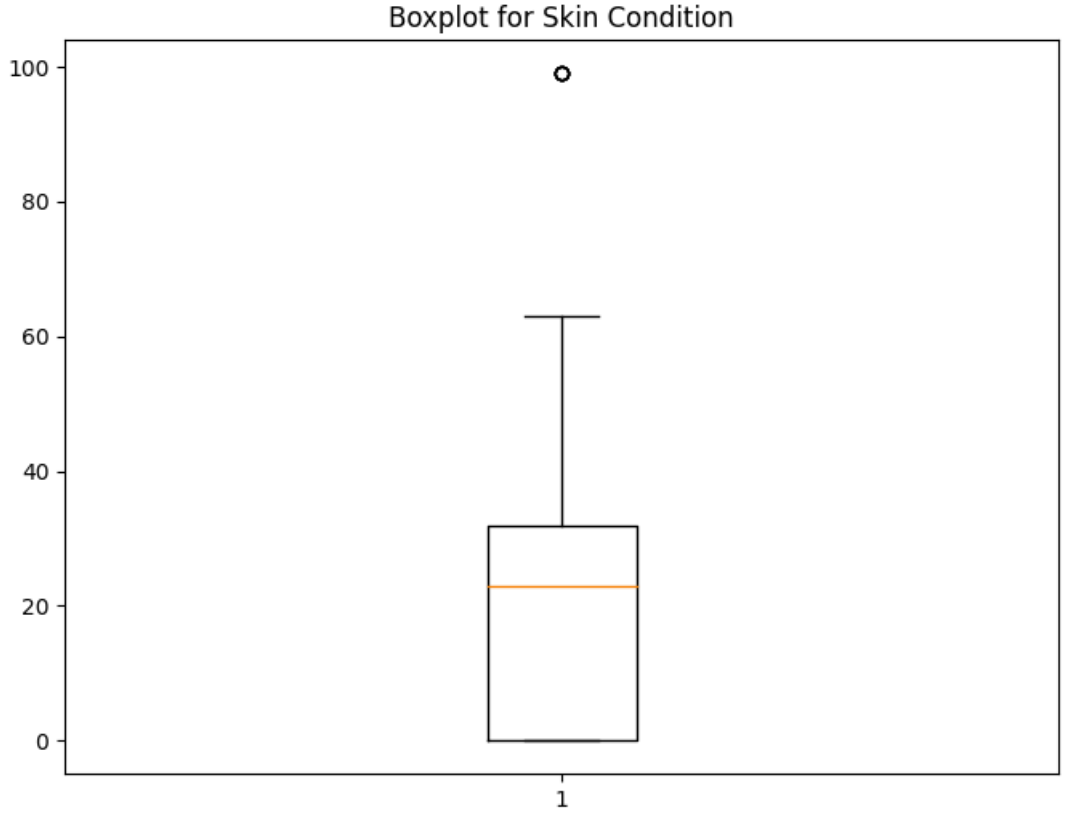
Now let's look deeper into particular variables. Here we present **histograms** for some sample variables...





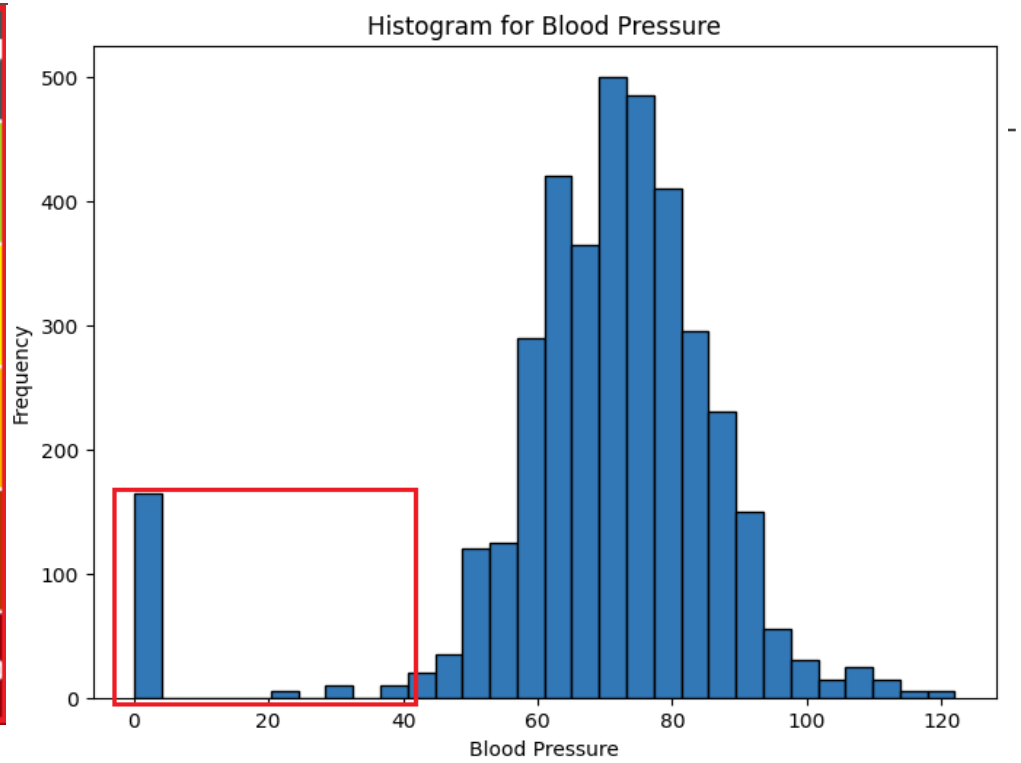
...and **boxplots** for the same sample variables.





Using these graphical representations and leverage well known knowledge we can notice some outlier values. Using commonly known information and information based on histogram we know that blood pressure indicate diastolic pressure and can take the values shown in the diagram below (on the left side):

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

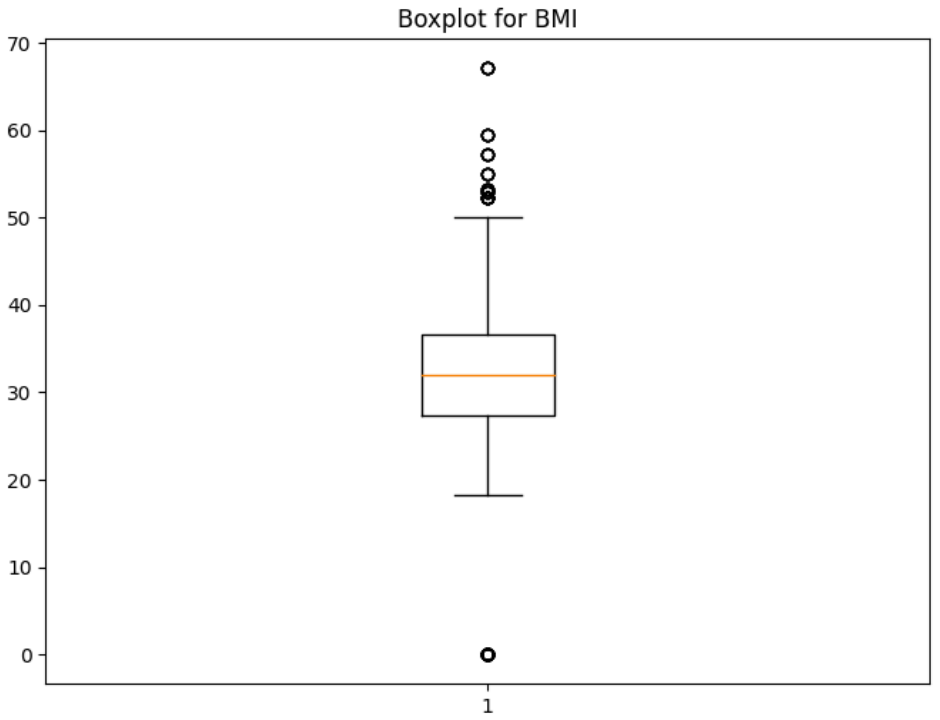


However histogram above shows us that the data consist some outlier values. We decided to set a outlier treshold on 40 or less. For all of outliers we change the value of blood pressure for the mean of blood pressure depending on what outcome record has. If a patient has a disease then we set a blood pressure value as a blood pressure mean for all patients with disease (outcome = 1). Otherwise we set blood pressure value as a blood pressure mean for all patients without disease (outcome = 0).

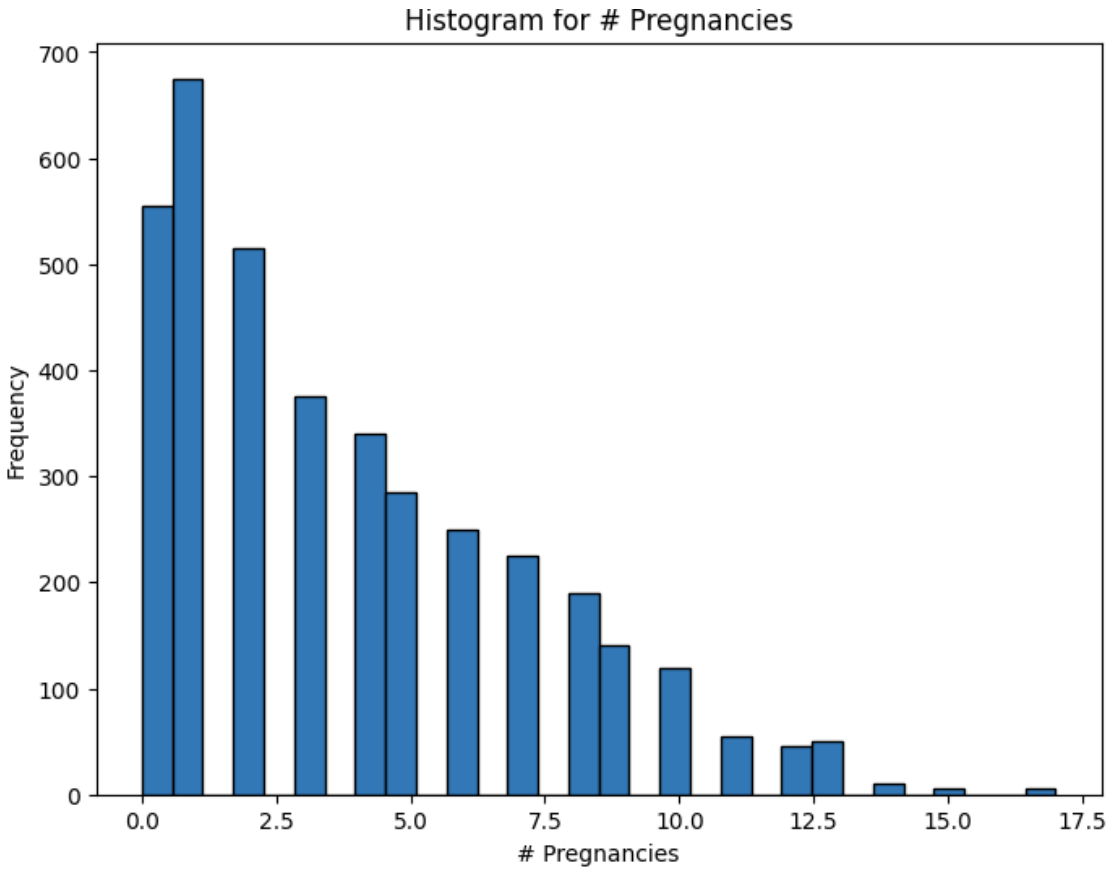
We decided to make a similar approach for BMI variable. Based on infromation in the table below...

HEIGHT		WEIGHT lbs	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320	330
in	cm	kgs	54.4	59.0	63.5	68.0	72.6	77.1	81.6	86.2	90.7	95.2	99.8	104.3	108.9	113.4	117.9	122.5	127.0	131.5	136.1	140.6	145.1	149.7
4'10"	124.5		25	27	29	31	34	36	38	40	42	44	46	48	50	52	54	57	59	61	63	65	67	69
4'11"	149.9		24	26	28	30	32	33	36	38	40	43	45	47	49	51	53	55	57	59	61	63	65	67
5'0"	152.4		23	25	27	29	31	32	35	37	39	41	43	45	47	49	51	53	55	57	59	61	63	65
5'1"	154.9		23	25	26	28	30	32	34	36	38	40	42	44	45	47	49	51	53	55	57	59	61	62
5'2"	157.5		22	24	25	27	29	31	33	35	37	38	40	42	44	46	48	49	51	53	55	57	59	60
5'3"	160.0		21	23	25	27	28	30	32	34	36	37	39	41	43	44	46	48	50	51	53	55	57	59
5'4"	162.6		21	22	24	26	28	29	31	33	34	36	38	40	41	43	45	46	48	50	52	53	55	57
5'5"	165.1		20	22	23	25	27	28	30	32	33	35	37	38	40	42	43	45	47	48	50	52	53	55
5'6"	167.6		19	21	23	24	26	27	29	31	32	34	36	37	39	40	42	44	45	47	49	50	52	53
5'7"	170.2		19	20	22	24	25	27	28	30	31	33	35	36	38	39	41	42	44	46	47	49	50	52
5'8"	172.7		18	20	21	23	24	26	27	29	30	32	34	35	37	38	40	41	43	44	46	47	49	50
5'9"	175.3		18	19	21	22	24	25	27	28	30	31	33	34	36	37	38	40	41	43	44	46	47	49
5'10"	177.8		17	19	20	22	23	24	26	27	29	30	32	33	35	36	37	39	40	42	43	45	46	47
5'11"	180.3		17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41	42	43	45	45
6'0"	182.9		16	18	19	20	22	23	24	26	27	29	30	31	33	34	35	37	38	39	41	43	43	45
6'1"	185.4		16	17	19	20	21	22	24	25	26	28	29	30	32	33	34	36	37	38	40	41	42	44
6'2"	188.0		15	17	18	19	21	22	23	24	26	27	28	30	31	32	33	35	36	37	39	40	41	42
6'3"	190.5		15	16	18	19	20	21	23	24	25	26	28	29	30	31	33	34	35	36	38	39	41	41
6'4"	193.0		15	16	17	18	20	21	22	23	24	26	27	28	29	30	32	33	34	35	37	38	39	40
6'5"	195.6		14	15	17	18	19	20	21	23	24	25	26	27	29	30	31	32	33	34	36	37	38	39

Underweight: < 18.5 | Healthy: 18.5 - 24.9 | Overweight: 25 - 29.9 | Obese: 30 - 39.9 | Severely Obese: ≥ 40



...and looking at boxplot we can assume that all values less than 15 and higher than 55 are outliers and we can change it to mean value based on outcome value (as in Blood Pressure).



Let's look at the distribution of the data on the number of pregnancies. We can see that despite a few larger values, we do not treat them as outliers - these values may be real. However, the number 0, indicating no pregnancies for the patient, attracts attention. Because of the lack of a variable indicating the patient's gender, the number 0 here may indicate a woman who did not get pregnant or a man. We must take this into account when making recommendations to the doctor on reporting future data.

We notice also some records with outstanding values such as 230-year-old patients.

	ID	# Pregnancies	BC I	BC II	BC III	Blood Pressure	Skin Condition	BMI	GPF	Age	Air Quality Index	State	Outcome
159	5642277	1.0	97.0	82.0	262.0	64.0	19.0	18.2	0.299	230.0	58.0	WI	0.0
916	5642277	1.0	97.0	82.0	158.0	64.0	19.0	18.2	0.299	230.0	58.0	WI	0.0
1673	5642277	1.0	97.0	82.0	55.0	64.0	19.0	18.2	0.299	230.0	58.0	WI	0.0
2430	5642277	1.0	97.0	82.0	205.0	64.0	19.0	18.2	0.299	230.0	58.0	WI	0.0
3187	5642277	1.0	97.0	82.0	307.0	64.0	19.0	18.2	0.299	230.0	58.0	WI	0.0

We decided to delete them from our dataset due to small volume of it. It was only 5 records so it should not influence on our dataset information.

After all the outlier manipulation is done, we can now deal with the missing values in our dataset. We do this step now so that the outliers that were in our dataset do not affect the values that we will replace the missing data with. All of the missing data are numeric values, so we will replace them similarly to some outliers in previous cases (e.g. Blood pressure and BMI) with a mean that depends on the values in the 'outcome' column.

```
ID                   0
# Pregnancies        0
BC I                25
BC II               10
BC III               0
Blood Pressure       0
Skin Condition       5
BMI                  0
GPF                  0
Age                  0
Air Quality Index   30
State                0
Outcome              0
dtype: int64
```

We noticed that some of the records have 'KU' abbreviation in State column. We assume that those were spelling mistakes and it means Kentucky ('KY'). We excluded 'ID' column from our dataset – it does not gives us any valuable information.

	ID	# Pregnancies	BC I	BC II	BC III	Blood Pressure	Skin Condition	BMI	GPF	Age	Air Quality Index	State	Outcome
12	5642130	2.0	68.0	66.0	49.0	70.0	32.0	25.0	0.187	25.0	0.0	KY	0.0
45	5642163	2.0	81.0	76.0	67.0	72.0	15.0	30.1	0.547	25.0	2.0	KY	0.0
101	5642219	1.0	90.0	59.0	89.0	62.0	18.0	25.1	1.268	25.0	90.0	KY	0.0
116	5642234	1.0	92.0	41.0	14.0	62.0	25.0	19.5	0.482	25.0	36.0	KY	0.0
139	5642257	1.0	95.0	38.0	53.0	66.0	13.0	19.6	0.334	25.0	52.0	KY	0.0
214	5642333	6.0	103.0	190.0	269.0	72.0	32.0	37.7	0.324	55.0	31.0	KY	0.0
288	5642407	1.0	112.0	176.0	316.0	72.0	30.0	34.4	0.528	25.0	60.0	KY	0.0
311	5642431	1.0	116.0	180.0	257.0	78.0	29.0	36.1	0.496	25.0	60.0	KY	0.0
512	5642638	10.0	101.0	0.0	60.0	86.0	37.0	45.6	1.136	38.0	78.0	KU	1.0
531	5642657	5.0	109.0	129.0	318.0	62.0	41.0	35.8	0.514	25.0	92.0	KY	1.0
542	5642668	9.0	112.0	0.0	22.0	82.0	24.0	28.2	1.282	50.0	24.0	KY	1.0
623	5642749	5.0	139.0	160.0	56.0	80.0	35.0	31.6	0.361	25.0	89.0	KU	1.0
656	5642782	12.0	151.0	271.0	15.0	70.0	40.0	41.8	0.742	38.0	32.0	KY	1.0

Now we can describe some descriptive statistics to our variables...

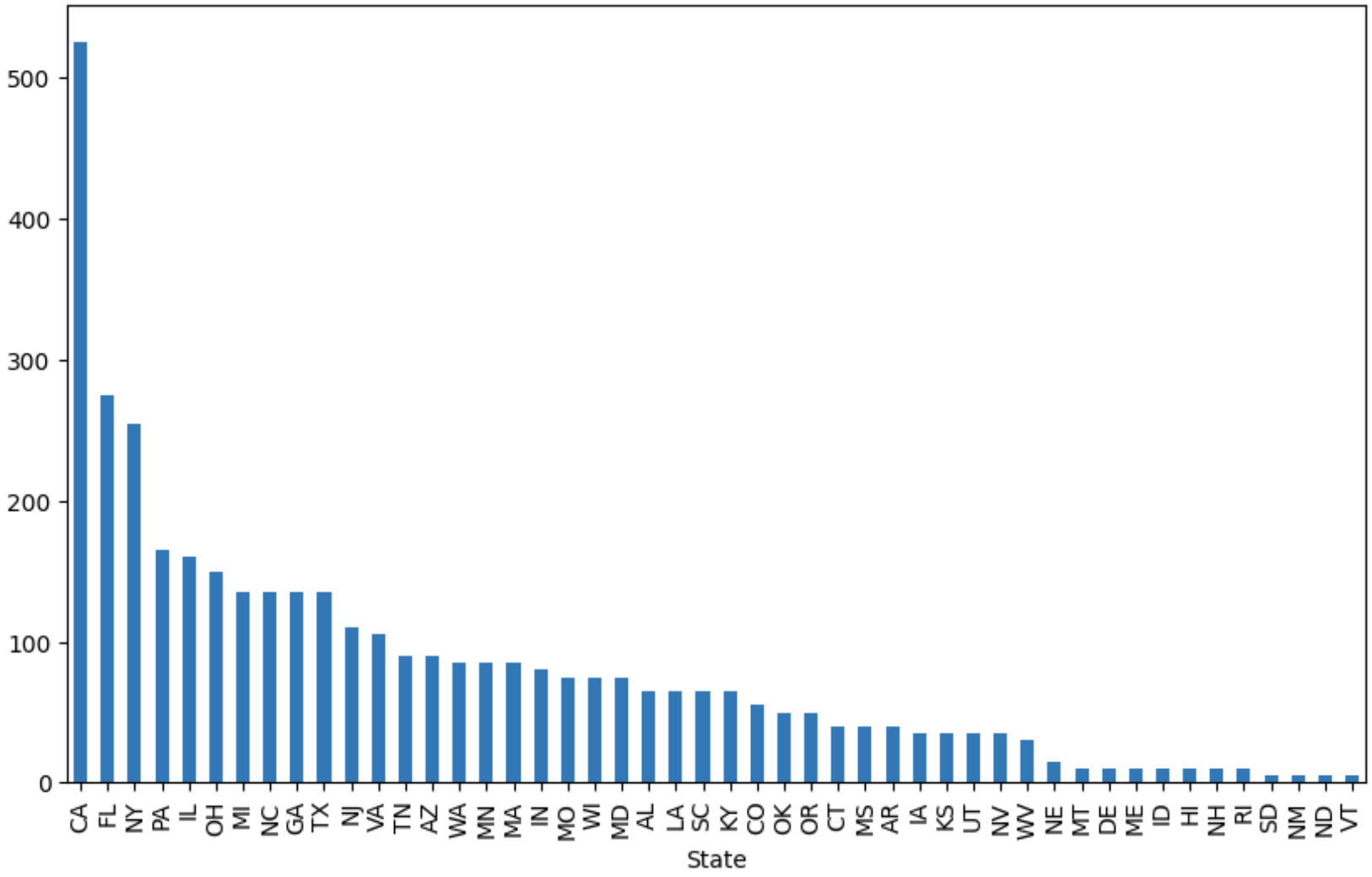
	# Pregnancies	BC I	BC II	BC III	Blood Pressure	Skin Condition	BMI	GPF	Age	Air Quality Index	Outcome
count	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000
mean	3.848761	120.918150	79.829304	174.134289	72.336211	20.543250	32.349299	0.472102	33.256845	49.095519	0.349413
std	3.368449	31.944008	115.202662	100.907732	11.988499	15.953389	6.617003	0.331313	11.753444	29.068206	0.476847
min	0.000000	0.000000	0.000000	0.000000	30.000000	0.000000	18.200000	0.078000	21.000000	0.000000	0.000000
25%	1.000000	99.000000	0.000000	86.000000	64.000000	0.000000	27.500000	0.243000	24.000000	24.000000	0.000000
50%	3.000000	117.000000	32.000000	175.000000	72.000000	23.000000	32.000000	0.374000	29.000000	49.000000	0.000000
75%	6.000000	141.000000	127.000000	259.000000	80.000000	32.000000	36.500000	0.627000	41.000000	72.000000	1.000000
max	17.000000	199.000000	846.000000	350.000000	122.000000	99.000000	55.000000	2.420000	81.000000	100.000000	1.000000

… and the same descriptive statistics after normalizing the data which we apply to avoid big influence of potential outliers which we did not manage to mitigate and have a possibility to compare variables to each other.

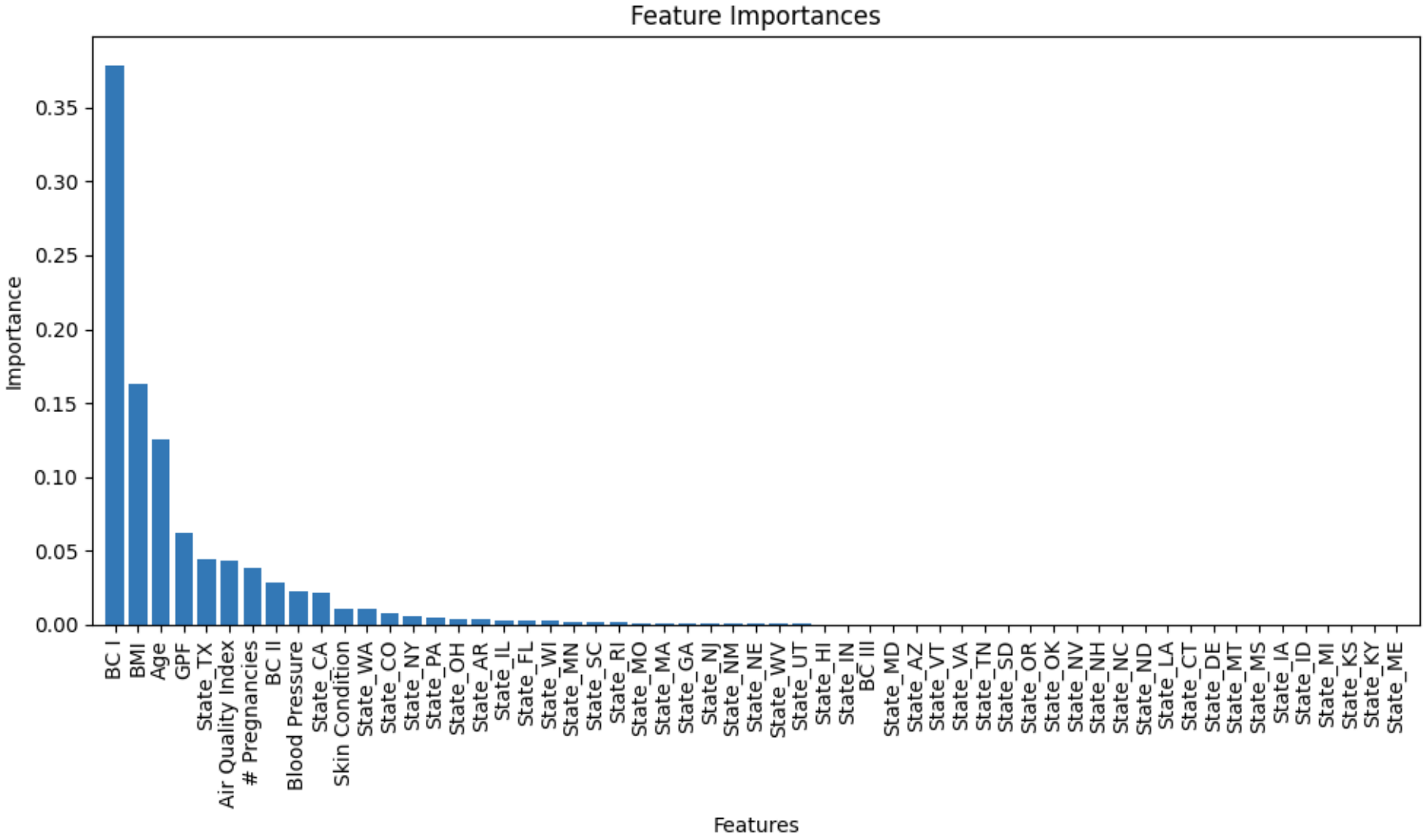
	# Pregnancies	BC I	BC II	BC III	Blood Pressure	Skin Condition	BMI	GPF	Age	Air Quality Index	Outcome
count	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000	3835.000000
mean	0.226398	0.607629	0.094361	0.497527	0.460176	0.207508	0.384492	0.168276	0.204281	0.490955	0.349413
std	0.198144	0.160523	0.136173	0.288308	0.130310	0.161145	0.179810	0.141466	0.195891	0.290682	0.476847
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.058824	0.497487	0.000000	0.245714	0.369565	0.000000	0.252717	0.070453	0.050000	0.240000	0.000000
50%	0.176471	0.587940	0.037825	0.500000	0.456522	0.232323	0.375000	0.126388	0.133333	0.490000	0.000000
75%	0.352941	0.708543	0.150118	0.740000	0.543478	0.323232	0.497283	0.234415	0.333333	0.720000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

We noticed that the blood chemistry type III and the air quality index had the largest deviations from the mean, as indicated by the high value of the standard deviation. It means that they are the most diverse. On the other hand, Blood Chemistry type II and Genetic Prediposition Factor are the variables most clustered to the mean, which means less variation in their measurements.

The distribution of data in terms of residence status is also valuable information. We can see it in the chart below.

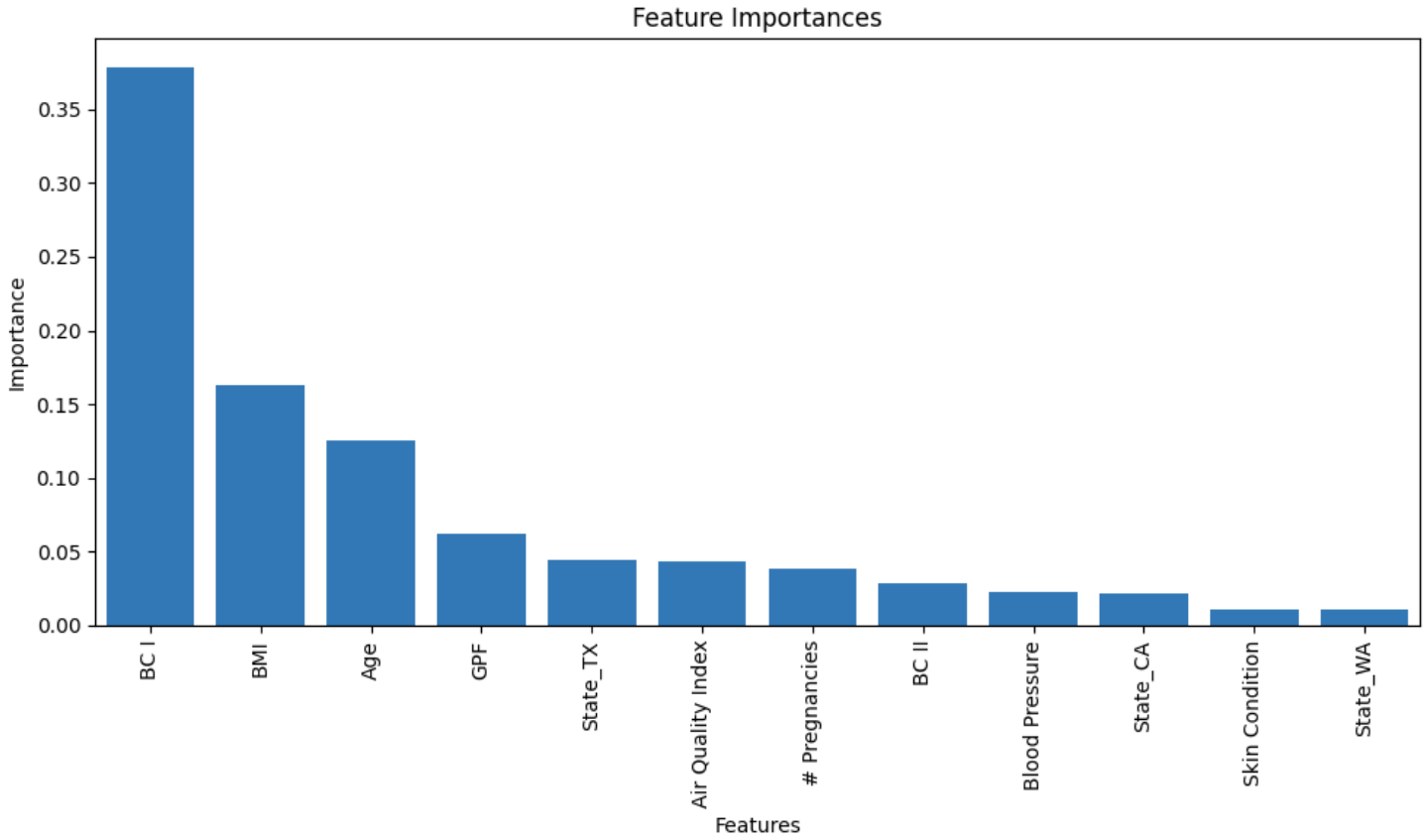


Let's look at some patterns which we notice in the data. The data consist a lot of variables so we want to check which of them have the greatest impact on the outcome of whether a given patient is sick or not.
Here we can see a feature importance ranking...

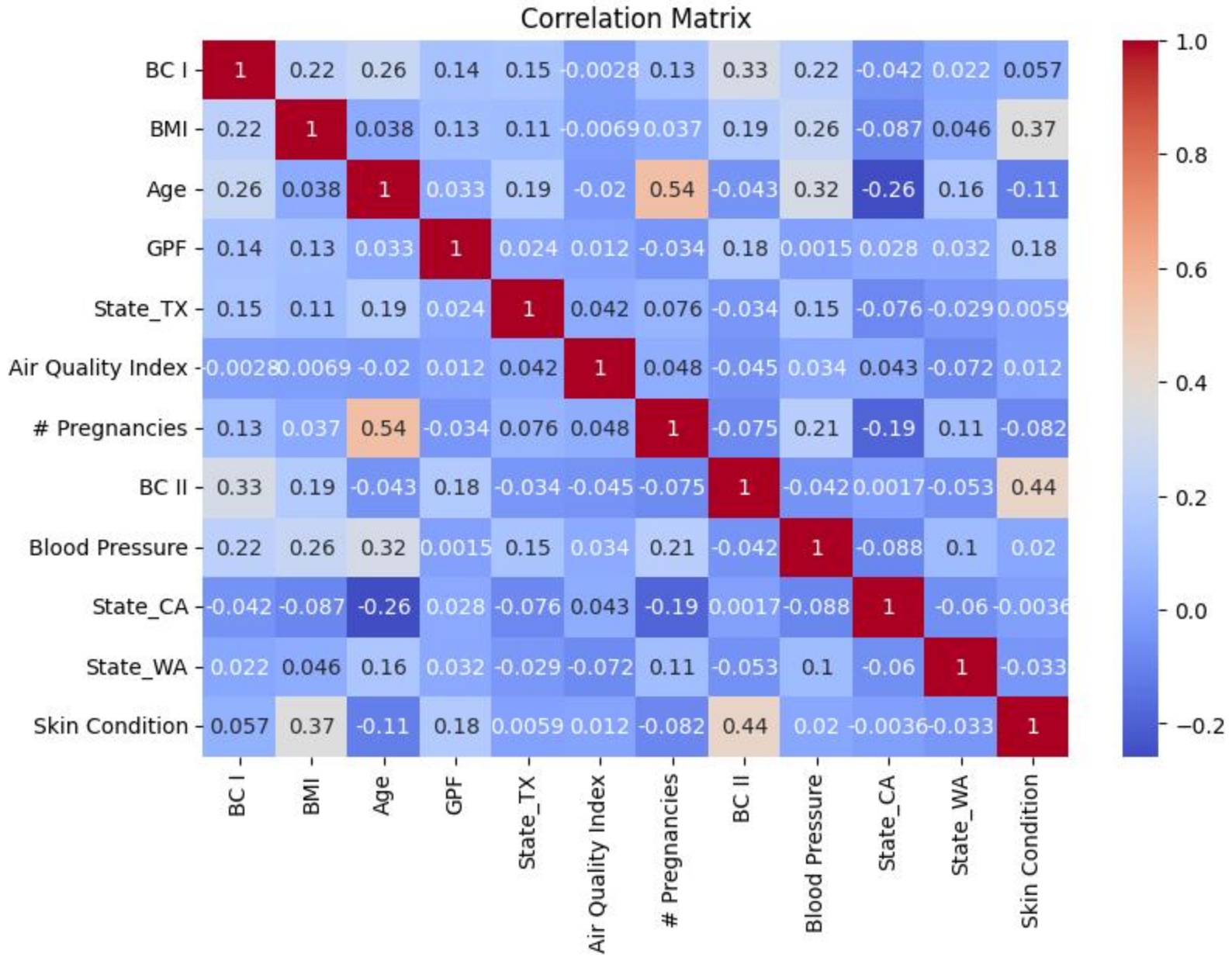


We focus on most impactful features

As we can see Blood Chemistry type I, BMI and Age are the most impactful variables on the result of the outcome. What is also interesting is that living in the states of TX, CA or WA was one of the most influential factors for the presence of the disease. We decided to take into account all features with at least 1% level of importances in building model process to avoid overfit of our models.



Let's look also at some correlations between variables which we can see in the data. We used the same features as we will use to build models so that it contains most influential features and make our chart more transparent and easy to read.



Here is some results that we can notice:

- the largest positive correlation is between age and the number of pregnancies (0.54)
- skin condition has high positive correlation rates between Blood Chemistry II and BMI (0.44 and 0.37, respectively)
- Blood Chemistry I and Blood Chemistry II are also related to each other with a high positive correlation (0.34)
- positive correlation index between age and blood pressure (0.32)
- positive correlation between BMI and blood pressure (0.26)
- negative correlation coefficient between age and residence in the state of CA (-0.26)

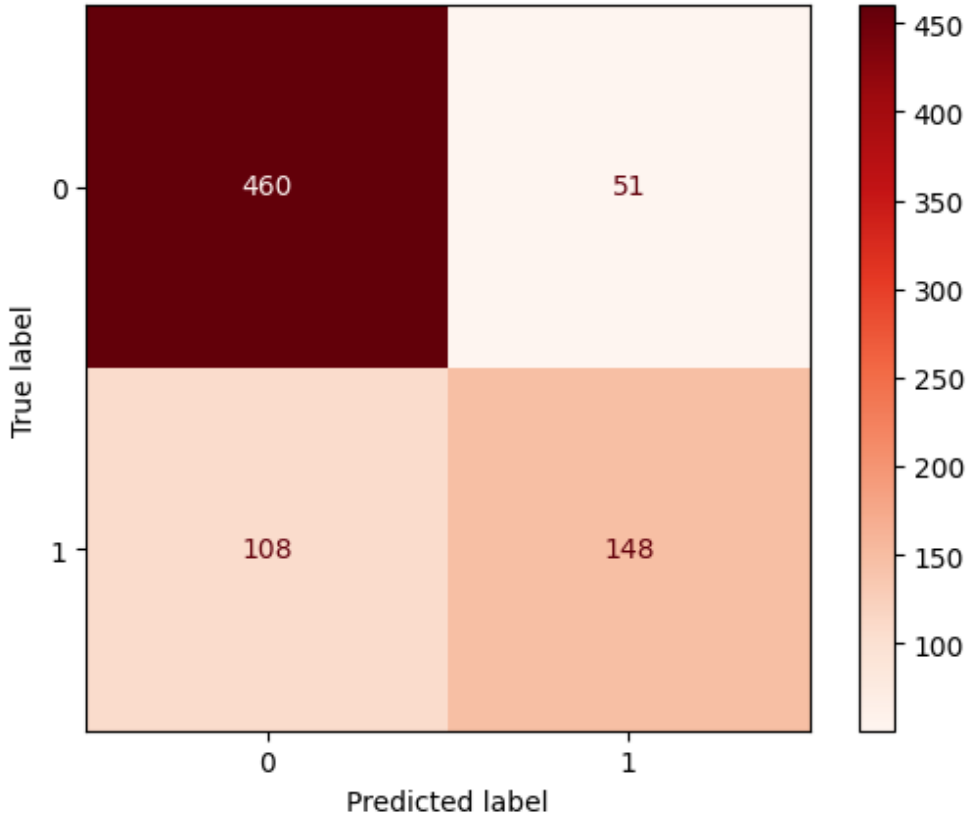
We can interpret them as follows:

- older patients may have more children
- poor skin condition may indicate a high BMI and poor blood test results
- a high Blood Chemistry type I rate may also indicate high Blood Chemistry type II rates
- with age, patients' blood pressure worsens
- patients with higher BMI may also have worse blood pressure
- patients residing in California are young people

We can confirm that California patients are young based on the average age in our database. We can also see what the average age is in the state of Texas. What's more interesting, investigating why the high importance value of residence in the state of Texas is, we can notice that out of 135 records from this state, 130 of them have a disease. This may be important information when ordering specific tests for people from the state of Texas.

```
Average age in CA: 25.666666666666668
Average age in other states: 34.4607250755287
Average age in TX : 44.81481481481482
Average age in other states: 32.83513513513513
Number of records from TX: 135
Number of records from TX with disease: 130.0
```

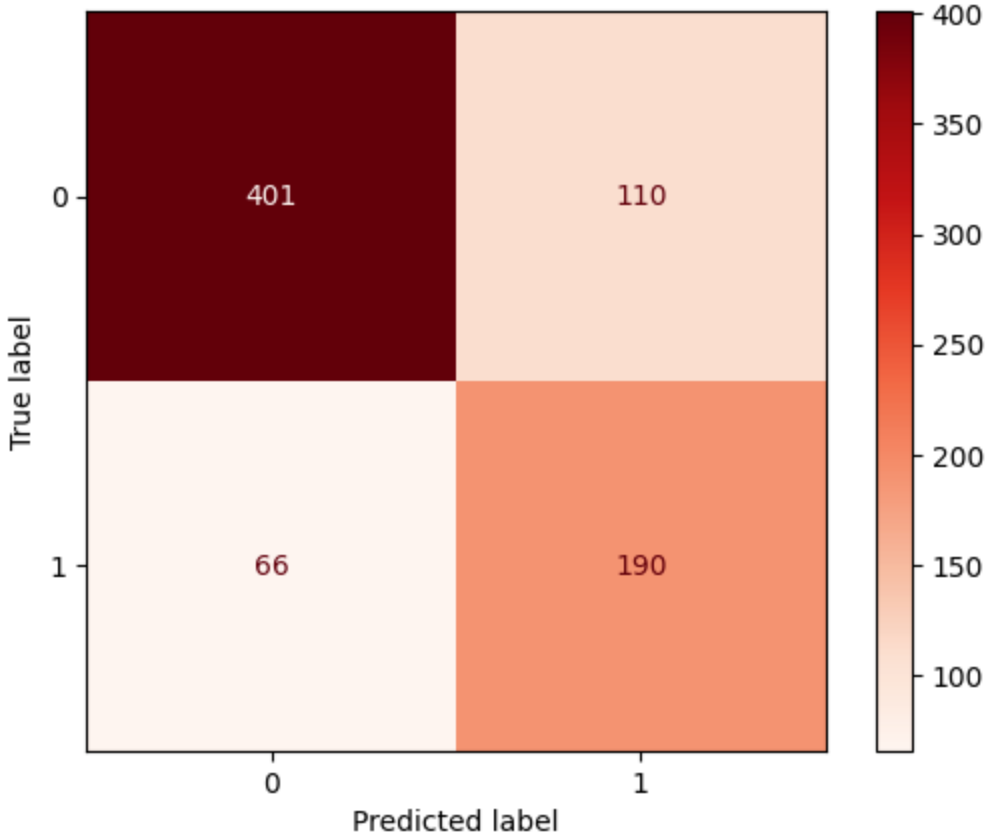
We also created 2 predictive models - using neural networks and logistic regression where we used 2 approaches: first we trained the model on imbalanced data that was provided to us by the doctor. Then we compared the same model - logistic regression - trained on balanced data using weights for individual classes.




```
Accuracy: 0.7926988265971316
Precision: 0.7437185929648241
Recall: 0.578125
F1 score: 0.6505494505494506
```

As we can see in the **logistic regression** model that used **imbalanced** data for training the prediction accuracy result is 79.2% which is a decent result. However, in case we are dealing with data where the minority class is much smaller than the majority class a better metric to evaluate the model would be the f-1 score – in that case we have 0.6505.

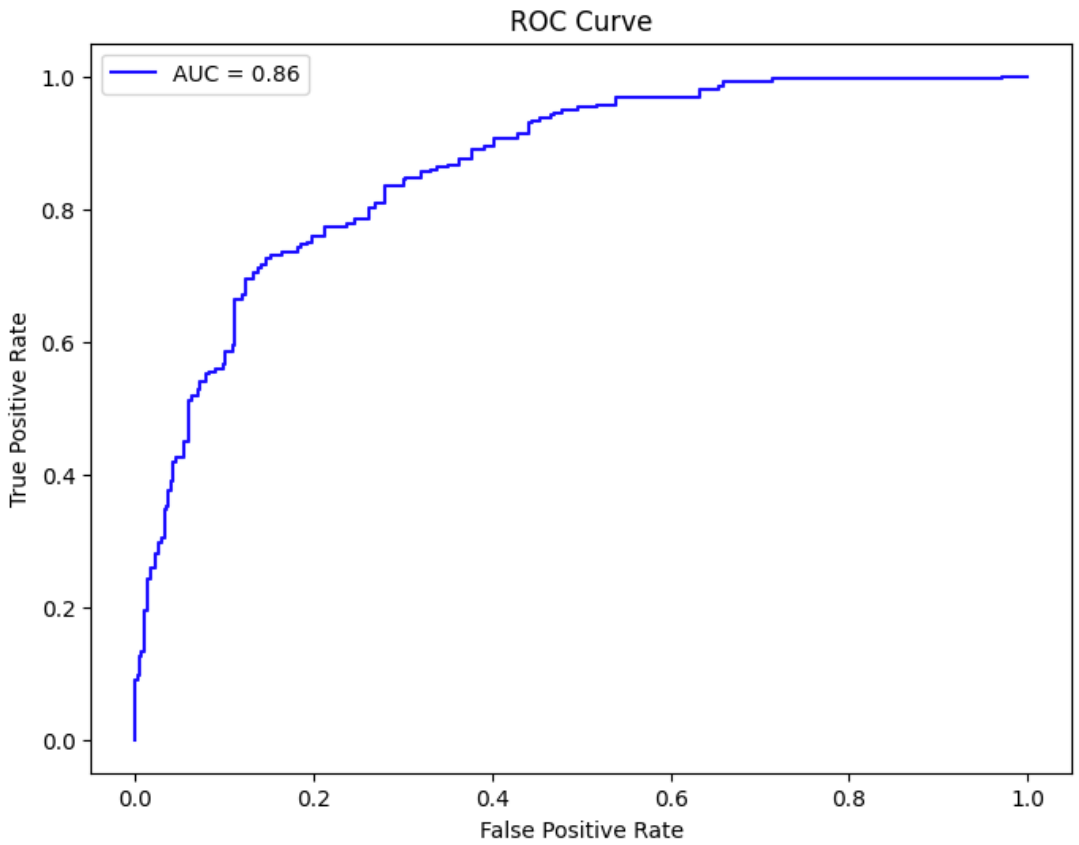
To apply specific weights to the **logistic regression** model with **balanced data**, we used the information contained in our data. We calculated the imbalance of our data (minority class/majority class) which was 0.54. We assigned this weight to class 0 and left the default weight - 1 for class '1' in the 'outcome' variable.

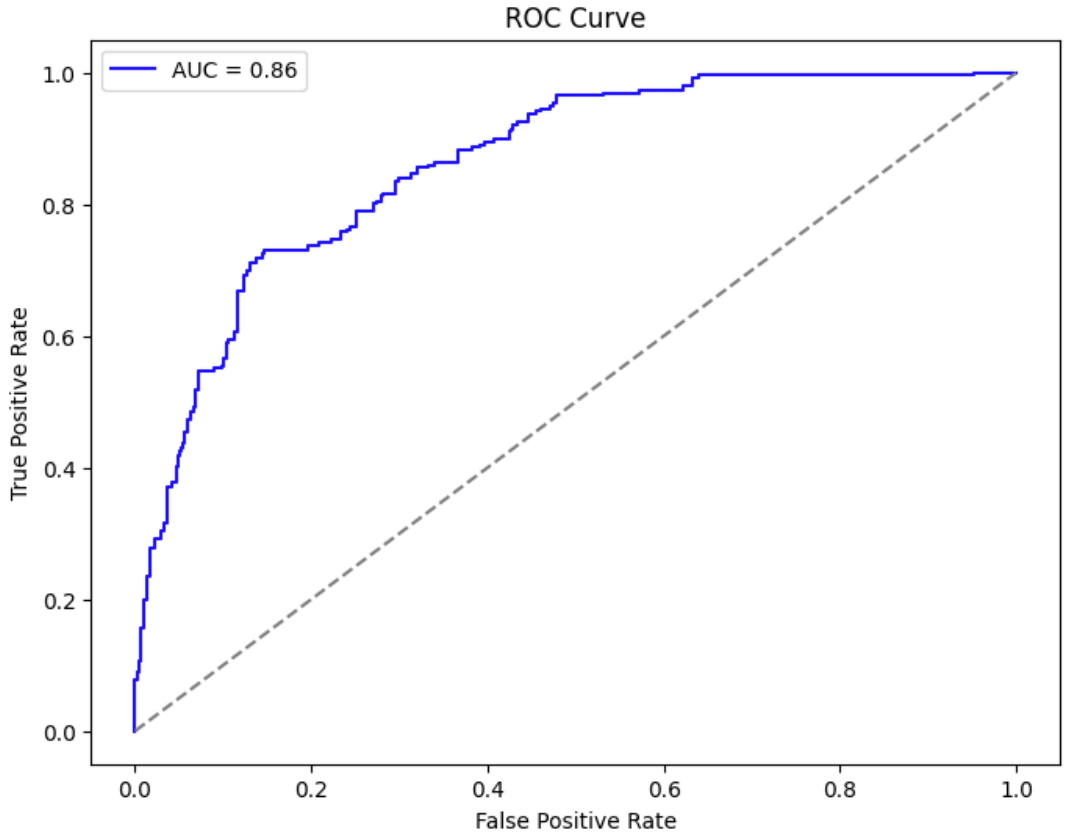



```
Accuracy: 0.7705345501955672
Precision: 0.6333333333333333
Recall: 0.7421875
F1 score: 0.6834532374100719
```

As we can see, this model has a slightly lower prediction accuracy - 77.05%, but it copes better with detecting patients with the disease, as evidenced by the recall result and the mentioned f1-score, which is 0.6834.

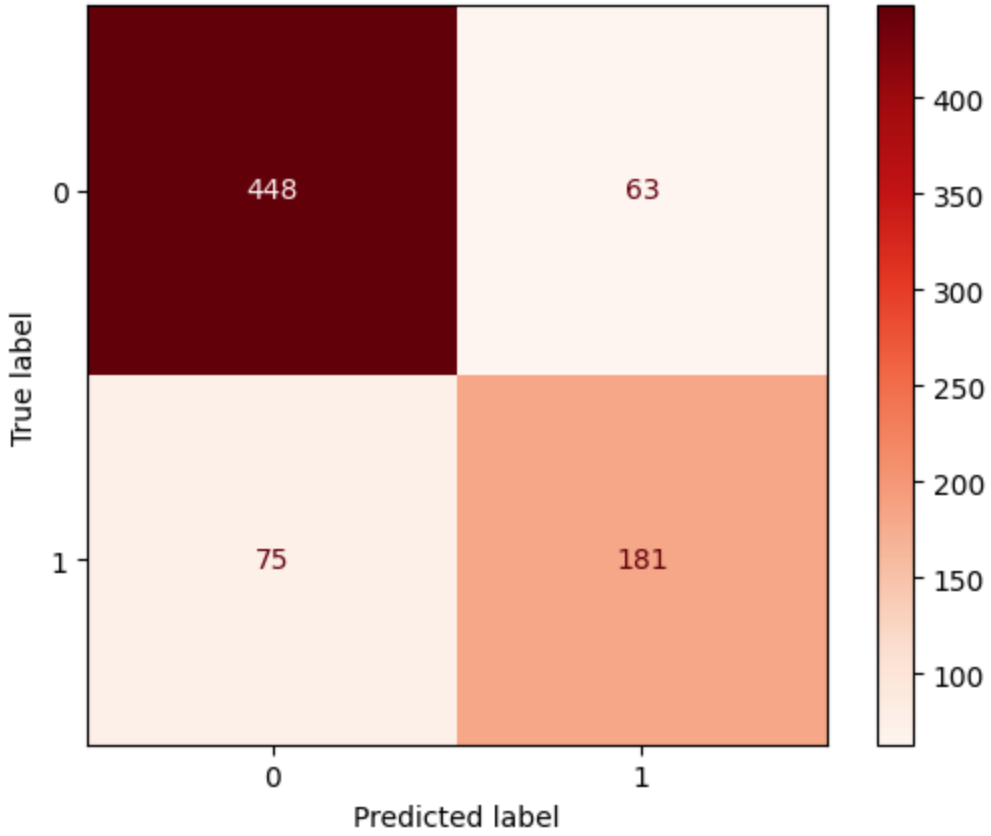
When it comes to evaluating both models using roc curve and AUC, the result is 0.86, which is a very good result and means that both models distinguish patients very well in terms of classes.





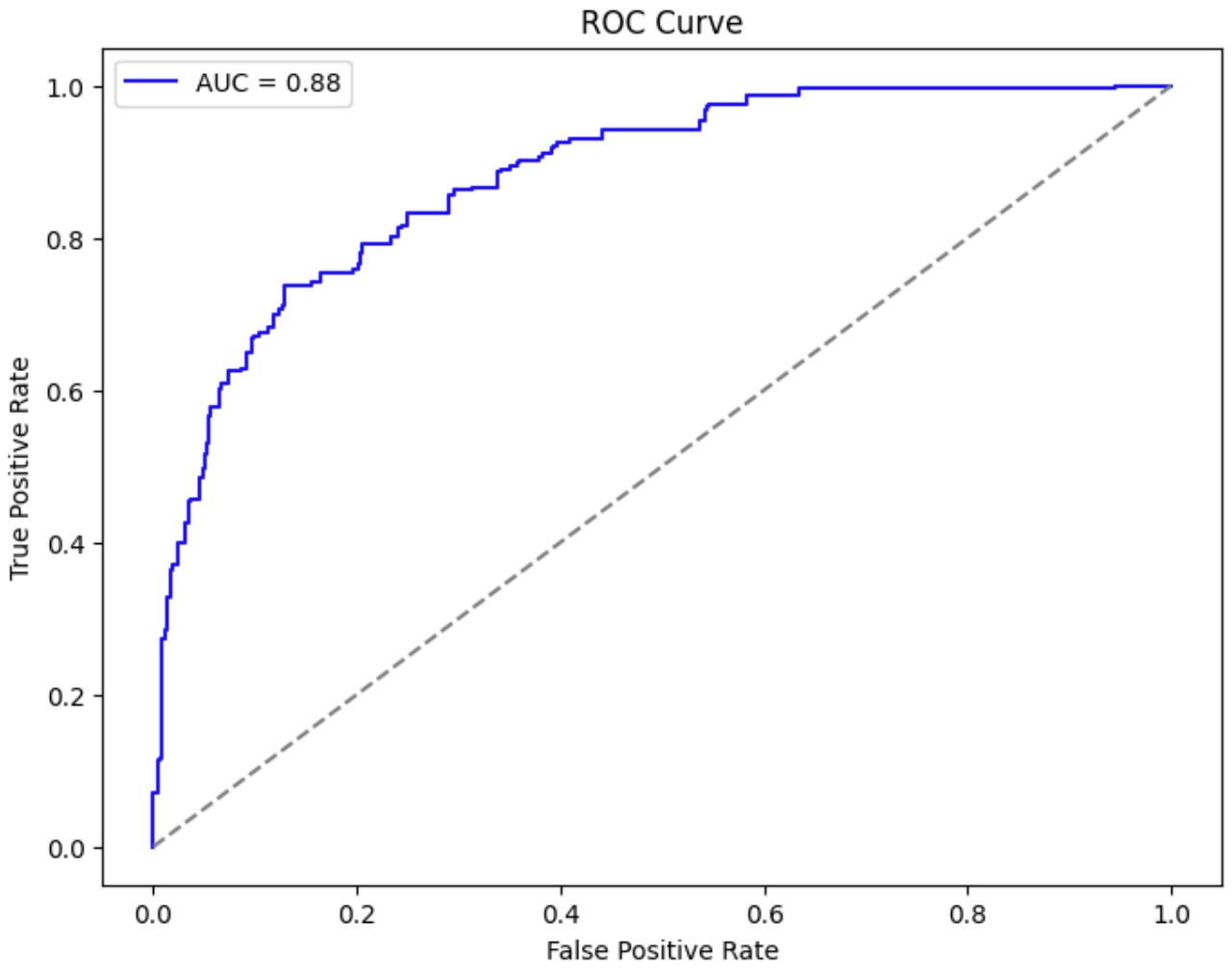
Now let's see how the results of the neural network model look like.

In the **neural network** model, we used the oversampling method to generate synthetic data for the minority class to remove the data imbalance problem. As we can see, this model performed the best of all three approaches, achieving 82% accuracy and an f1-score of 0.724.




```
Accuracy: 0.8200782268578879
Precision: 0.7418032786885246
Recall: 0.70703125
F1 score: 0.724
```

We also see that this model is even better at distinguishing patients with the disease from those without it.



Recommendations

- regular screening for Blood Chemistry type I and BMI,
- more frequent tests for older patients,
- due to the higher Genetic Predisposition Factor feature importance, monitoring family history related to the disease,
- detailed and more comprehensive tests for Texas patients,
- paying attention to the condition of patients' skin,
- more frequent blood pressure monitoring in older people,
- more attention when recording patient data,
- gender reporting to improve patient discrimination and future model improvements and also better recommendations based on patient gender

References:

- https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
- https://www.builtlean.com/bmi-chart/